European Journal for Philosophy of Science (2021) 11: 57
https://doi.org/10.1007/s13194-021-00377-2

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

How (not) to measure replication

Samuel C. Fletcher[1]

Received: 27 January 2020 / Accepted: 6 May 2021 / Published online: 3 June 2021

Abstract
The *replicability crisis* refers to the apparent failures to replicate both important and typical positive experimental claims in psychological science and biomedicine, failures which have gained increasing attention in the past decade. In order to provide evidence that there *is* a replicability crisis in the first place, scientists have developed various measures of replication that help quantify or "count" whether one study replicates another. In this nontechnical essay, I critically examine five types of replication measures used in the landmark article "Estimating the reproducibility of psychological science" (Open Science Collaboration, *Science, 349*, ac4716, 2015) based on the following techniques: subjective assessment, null hypothesis significance testing, comparing effect sizes, comparing the original effect size with the replication confidence interval, and meta-analysis. The first four, I argue, remain unsatisfactory for a variety of conceptual or formal reasons, even taking into account various improvements. By contrast, at least one version of the meta-analytic measure does not suffer from these problems. It differs from the others in rejecting dichotomous conclusions, the assumption that one study replicates another or not simpliciter. I defend it from other recent criticisms, concluding however that it is not a panacea for all the multifarious problems that the crisis has highlighted.

Keywords Replicability crisis · Reproducibility crisis · Null hypothesis significance testing · Effect size · Confidence interval · Meta-analysis

1 Introduction

Why run a seemingly successful scientific study again? Suppose that it found evidence for an interesting effect and even estimated that effect's strength or magnitude.

This article belongs to the Topical Collection: *Philosophical Perspectives on the Replicability Crisis*
Guest Editors: Mattia Andreoletti, Jan Sprenger

✉ Samuel C. Fletcher
scfletch@umn.edu

[1] Department of Philosophy & Minnesota Center for Philosophy of Science, University of Minnesota, Twin Cities, Minneapolis, MN, USA

If everything is as it seems with the study in question, it is highly likely that running it again "will successfully produce the same or sufficiently similar results as the original" (Fidler and Wilcox 2018, §1). The original such study would be called *replicable* and the new study its *replication*. Replicable studies are widely considered to be, at least in principle, the sina qua non basis of much of workaday scientific knowledge (Schmidt 2009; Romero 2019).[1] This is because the replicability of scientific results reflects a kind of reliability, an assurance that scientists can dependably build on those results in further inquiry.

Understandably, then, *failures* of replicability in social psychology (Klein et al. 2014; OSC 2015), cancer biology (Begley and Ellis 2012; Nosek and Errington 2017), and other social scientific scientific fields (Camerer et al. 2018) have renewed scientific attention to the efficacy of current research and unsettled some scientists' confidence that they can safely and cumulatively build on published findings. Such failures arise from carefully attempted replications that did not produce even sufficiently similar results as the original. Although opinions differ on the main causes of replicability failures—see, e.g., Fidler and Wilcox (2018, §2) and Romero (2019, §3) for reviews thereof—there is evidence that the vast majority of scientists consider it to constitute a *crisis* of sorts. For example, in an online survey of 1,576 scientists, the journal *Nature* found that 90% of respondents agreed that there is a "reproducibility crisis"; 52% described the crisis as "significant" and 38% as "slight" (Baker 2016).

Prior to diagnosing the sources of the problem, one must establish the existence and extent of the problem; conceptually prior even to that, one must adopt some method of *measuring* replication. While within the scientific literature on replication there has been much discussion of what *sorts* of scientific studies can count as potential replications (Fidler and Wilcox 2018, §1)—what it would mean to "run it again"—perhaps surprisingly there has been comparatively little discussion of the most appropriate ways to *measure* replication. Most have assumed that a replication measure must employ a definition of when a potential replication is a success or failure—or, perhaps more precisely, what it would mean for the results of the potential replication to be "the same or sufficiently similar."[2] Indeed, even under this assumption, while the aforementioned studies exhibiting replicability failures and many others concerned with replicability adopt various prima facie plausible measures of replication, they do not systematically explore arguments for or against them. Yet, each measure in general entails *different* conclusions regarding the extent of a study's replicability, thus the severity of the replication crisis.

Such a more systematic (and critical) exploration is the primary goal of this essay. In particular, I focus on the fives classes of replication measures discussed in the

[1]Depending on the discipline and context in which they are used, terms derived from "replicable" and "reproducible" can be synonymous or not (Fidler and Wilcox 2018, §1). In this essay, these terms will not mark distinct concepts, although I will attempt to use only terms derived from "replicable"; I am in particular only concerned here with the kind, adumbrated above, sometimes known as *direct* replication (Schmidt 2009). (See, e.g., Nosek and Errington (2020) or Machery (2020) for alternative definitions and typologies.) However, I am not concerned here so much with the minutiae of its definition as with the techniques for measuring it; see the end of this Section 1 for further remarks thereon.

[2]One of my conclusions, discussed in Section 8, will be that the dichotomous terms of "success" and "failure" are inapt for measuring replication.

seminal work on replication in psychology by the Open Science Collaboration (OSC 2015):

1. a subjective assessment by experts;
2. whether one arrives at the same conclusions in a significance test of the null hypothesis that there is no effect;
3. the results of a significance test of a difference in effect sizes between the original and attempted replication;
4. comparing the original effect sizes with confidence intervals from the attempted replication; and
5. statistical meta-analytic methods.

Each of these classes employs calculations or statistics that differently summarize what the "results" of a scientific study are. I describe how in the first subsection of each of Sections 3–7, respectively, after introducing the common framework of classical statistical testing and estimation that they presuppose in Section 2.[3] Although OSC of course employ them in the context of the replicability of studies in psychological science, they apply generally to any studies that use statistical testing in their determination of results.

Now, OSC themselves proclaim that none of these classes of replication measures is without problems and do provide some brief criticisms of each of them.[4] In the second subsection of each of Sections 3–6, I critique each of the first *four* classes of measures on various conceptual and formal grounds. My critique in each case extends beyond OSC's; I also consider various modifications and responses to these criticisms that, I argue, are ultimately unsuccessful.[5]

Then, I defend in Section 7.2 a certain version of the fifth class of measures, based on the ideas of statistical meta-analysis, from OSC's and other criticisms. In a word, these criticisms erroneously extend problems for applying meta-analytic techniques in other, more traditional contexts to their use for measuring replication. The version of the meta-analytic measure I endorse—what Braver et al. (2014) call "continuously cumulating meta-analysis"—also avoids my criticisms of the other replication measures I discuss. But I emphasize in the concluding Section 8 that even though meta-analytic measures of replication are free from the problems plaguing other measures, they are not a panacea for the current challenges of replicability. In particular, they do not immunize scientific inquiry from bias or questionable research practices, whose extent must be estimated and modeled within a meta-analysis to make its conclusions deriving from our total evidence more accurate. Nevertheless, as a method for measuring replication, meta-analysis suggests moving beyond dichotomous measures to quantifying how replications change our total evidence for hypotheses of interest.

[3]Others have suggested interpreting OSC in Bayesian terms (Etz and Vanderkerckhove 2016), which I will address in Section 4.2.

[4]Cf. their statements that "There is no single standard for evaluating replication success" (OSC 2015, p. 2) and "No single indicator sufficiently describes replication success, and the five indicators examined here are not the only ways to evaluate reproducibility" (OSC 2015, p. 6).

[5]To be clear, my critique's focus is the replication measures, not OSC's particular employment of them or their conclusions about psychological science.

Before continuing to review OSC's different measures of replication, I wish to circumscribe the scope of the present investigation. Not all studies are candidate for replicating any other: they have to be sufficiently similar in the relevant ways. How can one make precise this vague notion of a candidate for direct replication? All replication measures presuppose some practically workable answer to this question. However, as I described in footnote 1, I will not prosecute that question here. In the remainder of this essay I presuppose that this question has been answered in some satisfactory way and that the hypothetical replication studies under consideration are genuine candidates for replications of the hypothetical original studies. This is possible because essentially none of the details of such an answer play a role in assessing replication measures: a definition of replication screens the replication candidates, but once they are passed, the definition has little to do with whether those potential replications "successfully produce the same or sufficiently similar results as the original."[6]

2 Measures of replication

All of the measures of replication that OSC consider apply to scientific studies analyzed in the setting of classical statistical testing and estimation. In such studies and their attempted replications, scientists collect data from a population to find evidence for the existence of a certain effect and various magnitudes it could have. For example, scientists might be interested in replicating the effects of different culturing conditions on the efficiency of pluripotent stem cells to differentiate into certain other types of cells (Patel and Alahmad 2016). One hypothesis states that a certain culturing condition has no effect. Others state that the condition has a positive or negative effect, to some specific degree—the *effect size*. Each of them, perhaps with some auxiliary assumptions about the experiment, entails a probability distribution for the differentiation efficiency of each resulting cell type. In the parlance of classical statistics, each of these is a *simple statistical hypothesis* for the data.

To test a given simple statistical hypothesis in a study, one typically amalgamates the data into a *statistic*—a function of the data—that orders all possible data according to how unexpected or extreme they are if that simple statistical hypothesis were true. The exact nature of the statistic is not so important for present purposes, only that the original and potential replication each use the same statistic (or one that is mathematically equivalent). For example, the data might consist of the percentage yields of healthy cells of a certain type after applying the culturing condition to several samples of stem cells, and the statistic might be those yields' mean.[7] The data are extreme with respect to some hypothesis about the effect of the culturing

[6]One way of criticizing a replication effort thus is to argue that the claimed replication studies do *not* qualify as candidate direct replications, as Gilbert et al. (2016) did with OSC (but see Anderson et al. 2016 for a reply).

[7]Good experimental design in these sorts of contexts dictates the comparison of the effect of the culturing condition on yields with some sample of stem cells left untreated, the so-called *control* group. I have omitted these details, which do not make any difference to the present illustration of classical statistical testing and estimation, for simplicity of presentation.

condition to the extent that the observed mean differs from the most likely mean, assuming that hypothesis were true. Given the mean of the data, one can compute that statistic's *p-value*: the probability of obtaining data *more* extreme than that actually collected, again assuming that hypothesis were true. The smaller the p-value, the more unlikely or incompatible the observed data are with the hypothesis, i.e., the stronger *evidence* the data provide *against* that hypothesis.[8]

With these concepts and techniques, the following three activities are of primary importance:

Null hypothesis significance testing One calculates the p-value of the statistic based on the hypothesis that there is an effect of a particular size, called the *null hypothesis*. If the p-value falls below some conventional threshold called the *significance level*—often 0.05 in psychology—then the data are understood to provide sufficient evidence against the null hypothesis to reject it. When a statistic is used for such a *significance test* in this way, it is called a *test statistic*. Often, the null hypothesis is that there is no effect—i.e., that the effect size is zero. In this case, the rejection of the null hypothesis entails the acceptance of the *existence* of the effect in question.

Point estimation of effect size Merely finding evidence for the existence of an effect does not entail anything about the size of that effect. One can use the value of a statistic based on the data, such as a mean as mentioned above, to estimate the actual effect size. Often, this is the effect size that maximizes the likelihood of the observed statistic or minimizes a function of the expected estimation error. When a statistic is used in this way, it is called an *estimator*, and the result a *point estimate*. As before, the particular nature and details of the estimator are not so important for present purposes, only that the original study and its replication use the same estimator.

Constructing confidence intervals for effect size Statistical hypotheses typically do not assign 0/1, or trivial, probability distributions to possible data: they assign probabilities (or probability densities) strictly between 0 and 1 to much possible data. Thus various possible data sets can be realized in an experiment, some more misleading about the true effect size than others. This affects the variability of point estimates in the same way. One way of representing this variability is not by estimating the effect size with a single number, but with an interval of numbers, called a *confidence interval*. Like with point estimates, confidence intervals depend on the data, but like with significance tests, these intervals depend on a conventional parameter called the *confidence level*—often 0.95—which gives the probability that the interval assignment procedure produces an interval containing the true effect size, assuming the adequacy of the modeling assumptions.[9] (The size of the interval tends to covary with the size of the confidence level).

[8]The way it does so is related to the concept of adherence in reliabilist epistemology (Nozick 1981). For more on concepts of evidence in classical statistics, see Fletcher and Mayo-Wilson (2021).

[9]It's important to remind ourselves that this probability is not that for any particular interval so produced to contain the true effect size, as would be for a Bayesian posterior interval. In classical statistical testing, statements like that are not even elements of the event space.

There is also an important connection between confidence intervals and significance tests. Given a test statistic calculated for a particular data set, and a particular significance level, one can ask: which (point) null hypotheses would a test of significance reject at that significance level? The complement of this set in the hypothesis space—all and only the effect sizes that the test would *not* reject—is a confidence interval with confidence level equal to one minus the significance level (Cox and Hinkley 1974, Ch. 7.2.iii). (This is why the commonly chosen confidence level of 0.95 pairs with the commonly chosen significance level of 0.05.) In other words, the effect sizes within a particular confidence interval represent those that are not so incompatible with the data as to necessitate rejection in a significance test.

With these three activities in mind, suppose that one is given an original study and an attempted replication thereof that have tested the same null hypothesis and each has produced a point estimate of the effect size and perhaps also a confidence interval for that effect. How does one measure whether (or the extent to which) the second study replicates the first? In each of the following sections I first describe and motivate the five classes of replication criteria that OSC use to answer this question. The goal is to show how each has at least some prima facie plausibility. Then I criticize the viability of the first four classes of measures and defend the viability of a version of the last, meta-analytic measure. The themes of these criticisms cluster around the unchecked possibility of objectionable bias (Section 3.2), the necessary reference to parameters whose values makes a difference to the replication measure but the selection of which is arbitrary or not otherwise grounded in reasons pertinent to the methodology, and asymmetries in how well the measures determine their outputs—replication successes and failures (Sections 4–6). The second theme is based on a commitment to the grounding of scientific evidential decisions in epistemically pertinent reasons, rather than the idiosyncratic and potentially biased judgments of individual scientists. The third is based on plausible formal properties of desirable replication measures: they should render verdicts about replication success and failure about equally well, and they should respect the symmetry of replication when modeled as a binary relation on studies representing "having the same or sufficiently similar results". The former bears on a replication measure's ability to support its function of helping to assess the reliability of scientific findings, and the latter bears on its fit with the formal properties desired of it as a relation between studies. These criticisms are thus not of the three activities mentioned above—significance testing, point estimation, or confidence intervals—but only of the particular replication measures that employ them.

3 Subjective assessment

3.1 Subjective assessment: description

One very simple method for assessing whether an attempted replication study produced the same or sufficiently similar results as an original is simply to ask the

replication team to answer the question, "Did your results replicate the original effect?" This is one of the replication measures used by OSC. Alternately, one might ask instead the team of the original study whether the new study replicates the results of the original.

OSC motivate this type of measure by the complexity of certain experimental designs and statistical models: "For more complex designs, ... quantitative analysis may not provide a simple interpretation," (2015 p. 4) while the scientific team may use their professional judgment to cut through this complexity. Furthermore, insofar as there is a fundamental sense in which scientific reasoning depends on human judgment, assessments of replicability ultimately depend on this judgment, too.

3.2 Subjective assessment: critique

There has been little discussion of subjective assessment as an appropriate replication measure, except perhaps insofar as it serves as a proxy for other, typically more formal replication measures (Dreber et al. 2015). In such proxy cases, scientists are given instructions describing a particular, more formal replication measure—such as that for null hypothesis significance testing in Section 4—and asked to judge whether a given study replicates the results of another, or predict whether the faithful implementation of a certain study design will replicate the results of another. (I will return to a discussion of these cases later in this section). By contrast, the subjective assessment measure that OSC invoked gives no instructions to interpret the question "Did your results replicate the original effect?" in any particular way. Scientists were free to interpret it as they see fit.

A first concern relevant to this (non-proxy) measure is that the *reasons for* a judgment about replication should be just as important as the judgment itself. Those reasons should be aligned with some theoretical account of what a replication is, so that human judgment is a reliable indicator of how its target would be classified on that account. Moreover, these reasons should track facts about the outcomes of experiments that are not themselves determined by or reducible to the outcomes of human judgment. Indeed, if one understands human bias as deviation from the facts or from an ideal scientific consensus determined (at least in part) from the facts, then the very possibility of human bias entails that those judgments do not completely determine or reduce to those facts.[10]

An analogy may illuminate this point. The sense in which I am claiming that the facts of replication are not determined by or reducible to human judgment is the same sense in which the results of an election are not, but by the votes cast and the election format. Someone's judgment (even repeated affirmation!) about the results of an election is biased when their method of judgment is not reliably determined by

[10]This entailment would not obtain if one understood human bias as deviation from scientific consensus *regardless of how that consensus was reached*. It is an assumption, albeit one that seems to be widely held and sometimes only implicitly in the scientific and philosophical literature concerning objectivity, replicability, and reproducibility, that a scientific community's methods for reaching consensus make a difference to whether those methods are objective. See, for example, Reiss and Sprenger (2017 §§4–5) and references therein.

and indicative of the facts about the votes cast and the election format. Such a method of judgment does not track and inform decisions about election outcomes that should depend on the relevant facts about the votes and format.

If the above concern is warranted, then it forms the crux of an objection to (non-proxy) subjective assessment as a replication measure: in its application, all sorts of irrelevant and biasing factors can influence the outcome of that measure. To the extent that these factors *could* have influence, the measure could be itself objectionably biased. What's to stop iconoclastic, dissentious researchers from making the judgment that they think will cause division? What's to stop defensive, careerist researchers from making the judgment that they think will advance their career? Division and career advancement are not always aligned with the epistemic goals of science. At the extreme end of this range, when the measure is given by the bare outcomes of human judgments alone, there is nothing to preclude or mitigate these biases. Adopting a purely subjective measure in this sense therefore seems to conflate the behavioral features of scientific judgment with their praxeological ones—the reasons in which scientific judgments are grounded.

One possible response to this objection invokes OSC's positive point about expert scientific judgment, that scientists may be able to assess the replicability of complex designs when formal criteria are hard to apply. If this expert judgment is known to be reliable, then it could suffice to measure replication absent any explicit theory or reasoning invoked for a given judgment. Empirical research on expert judgment, however, circumscribes the specific and sometimes narrow contexts in which it can be reliable (Shanteau 1992; Larrick and Feiler 2015). Crucially, "expert knowledge is acquired from experience and training. . . . For expertise to arise from experience and training, decision makers must be exposed to experiences that provide immediate, accurate feedback about relationships in the world" (Larrick and Feiler 2015, p. 697). For example, the skies provide immediate, quantifiable feedback to weather forecasters, the survival of patients to surgeons, etc. Each must have received this feedback from hundreds or thousands of cases before becoming reliable experts. By contrast, (non-proxy) assessments of replication are rare enough and provide little or no feedback about whether that assessment was correct. The number of replications most scientists assess is unknown, but is likely small due to lack of professional incentives for doing so (Romero 2017).

It's important to distinguish this lack of expertise for non-proxy subjective assessment measures with the substantial expertise many scientists have with proxy subjective assessment measures. For the latter measures, scientists would be given explicit formal criteria to apply in their judgment, such as those described in Sections 4–7. Scientists' experience applying these methods in their own work and being checked in these applications by their peers gives reason to believe in the reliability of their expertise in proxy mesaures. I am not aware of any replication efforts employing proxy subjective assessment measures, but they are popular in the related but still distinct task of *predicting* whether the results of a study will replicate in a prospective study with a specified experimental design (Dreber et al. 2015; Camerer et al. 2018; Forsell et al. 2019). In these tasks, scientists are provided, e.g., with the null hypothesis significance testing measure discussed in Section 4 and a number of trading resources that they can use to buy and sell contracts that pay out just

when a specific study's results replicate.[11] Their degree of belief in each such event is also solicited. Standard techniques in economics then allow one to reconstruct the collective ("market") probability of replication (Wolfers and Zitzewitz 2006). Success at this prediction task shows good evidence that scientists have expertise that cuts through the complexity of individual experiments, but only towards the explicit formal replication criterion used.

The aforementioned research on expert judgment suggests that experts rely on heuristics for their judgments just like everyone else, only that the experts' are trained from the whip of experience (Kahneman and Klein 2009). When their judgments are elicited, they may rely on those heuristics, even if the question asked did not prompt them directly. It is thus not so unexpected that in the OSC study, where the non-proxy replication measure was used, there was a substantial *correlation* between replication teams' subjective judgments of replication and the null hypothesis significance testing criteria described in Section 4 (OSC 2015, pp. 3–5). Plausibly, many interpreted the replication question that OSC asked as a proxy for assessing that criterion. If this interpretation were confirmed and widespread, then the subjective judgment replication measure would closely track the significance testing criterion because the scientists surveyed would be employing essentially the same reasoning for that criterion. But this crypto-proxy subjective measure of replication would then be vulnerable to analogous criticism, which I describe in Section 4.2, of the reasoning behind the criterion it proxies. In particular, there is reason to believe that it is not an unbiased measure of replication.

One possible further response to these difficulties would locate them in the imperfections of scientists as rational agents. Instead of determining replication through elicited expert judgment, it would model how scientists *ought* to judge as ideal Bayesian agents. For instance, Earp and Trafimow (2015) have developed a Bayesian framework that offers an explanation of the epistemic significance of replication attempts in terms of the ways they can modify the confirmation of the original study's results. However, in order to implement this idea, they *presuppose* that success or failure of replication is an event in the ideal Bayesian agent's probability space. That's to say that they take as given the existence of an unambiguous, dichotomous replication measure; for present purposes, *what that measure is* is precisely the issue in question.[12]

4 Null hypothesis significance testing

4.1 Null hypothesis significance testing: description

Another straightforward way to measure replication is to compare the results of null hypothesis significance tests (NHST) of the same hypothesis from both the original

[11] Forsell et al. (2019) also employ an effect size comparison measure like those discussed in Section 5.

[12] This is not a criticism of their project. Again, they aim to account for the significance of replication successes and failures, not how one measures replication.

and the attempted replication. Given a fixed significance level, one can classify both the original study and the attempted replication dichotomously according to whether they reject the null hypothesis (i.e., whether the p-value of the test fell below the significance level). The replication is successful if and only if the original study and attempted replication are accordingly classified in the same way and the point estimates for both have the same sign. (Usually a study that is the target of a replication effort has rejected a null hypothesis of interest, so an attempted replication would be successful when it rejects that hypothesis, too).

A rationale for this replication measure arises from the centrality of NHST in psychology, biology, and other disciplines. Psychological theories, for instance, typically do not offer quantitative predictions of effect sizes, but only the existence of an effect and the sign of the effect size—i.e., an inhibitory (negative) or promoting (positive) effect. Thus two studies that reject the null hypothesis of zero effect size with a point estimate of the same sign are both evidence for the same qualitative hypothesis.

4.2 Null hypothesis significance testing: critique

Even though NHST is ubiquitous in the social and biological sciences, criticism of NHST is (almost) just as ubiquitous in the methodology literature (Morrison et al. 1970; Harlow et al. 1997; Kline 2004; McCloskey and Ziliak 2008). NHST has also been identified as a possible cause for the replication crisis—see Romero (2019, p. 4) and Fidler and Wilcox (2018, §2.4) for reviews. Indeed, some of the criticisms of NHST are apropos to its use as a replication measure. In this subsection I focus on two such problems leveled at NHST: the misleading arbitrariness of its dichotomous division of outcomes, and its evidential asymmetry between rejection and non-rejection. At the end of the section, I suggest an underlying reason why NHST-based measures are vulnerable to these criticisms: NHST does not fully represent the *results* of a study, so focusing on it only cannot in general provide an adequate basis for judging replication.

Rosnow and Rosenthal (1989, p. 1277) summarize a common complaint about NHST: "surely God loves the .06 nearly as much as the .05." When the significance level is set to 0.05, NHST treats this "threshold as a bright-line criterion between replication success and failure" (OSC 2015, p. 4) so when the p-value of a test falls slightly above this—e.g., at 0.06—it entails a different conclusion, non-rejection, than if it were to fall equally slightly below. Yet these two p-values seem to represent quite similar evidence against the null hypothesis. Thus NHST is too coarse-grained to account for the similarity of studies with these p-values, hence it erroneously does not count one as a successful replication of the other.

Although ultimately I agree with the thrust and conclusion of this criticism, the NHST advocate has grounds for defense. If one takes the *results* of a study to consist solely in the conclusion about rejection that NHST offers, then this criticism begs the question by admitting the fine-grained p-values *also* as results of a study.[13] The

[13] To be clear, OSC do not take this position; as far as I know, it is a novel, if extreme, way of defending the use of NHST.

fact that a small change in a quantitative property (the p-value) can produce a change in a qualitative property (replication) should be no more troubling than other cases of linguistic vagueness and sorites paradoxes. Here, it seems most appropriate to employ a contextualist epistemic theory of vagueness, analogous to that of Graff Fara (Graff 2000; Graff Fara 2008), in which the extensions of vague predicates are relative to the current interests of a speaker.[14] With respect to the "rejection" concept, these interests determine the appropriate significance level below which the p-values warrant rejection, in accordance with some recent recommendations to researchers to justify the particular significance level they adopt in any particular study (Lakens et al. 2018). On this view, the significance level is not a convention so much as an expression or distillation of an individual researcher's views about what the results of a study are and what it would take to replicate them.

In practice, though, it is quite difficult to set such significance levels because all the considerations that would bear on them are rarely transparent to any researcher. Indeed, I know of no systemic account of how this should be done in general.[15] This is partly because it is unclear what epistemic or practical goal, exactly, is being considered. Without this transparency, the choice of significance level becomes a matter of judgment possibly subject to the same sorts of unchecked biases that non-proxy subjective judgment has as a replication measure (as discussed in Section 3.2). Since the exact value of the significance level makes a difference to the verdicts that the NHST replication criterion renders, these verdicts can therefore be objectionably biased. Setting the significance level as a convention may preclude such biases, but renders the facts of replication objectionably conventional, in that they are not appropriately grounded in the facts about the data and relevant goals of the research.

The second problem with using NHST as the basis for a measure of replicability is that it does not equally reliably indicate failures of replication as successes. The concept of the *power* of a statistical test illustrates why. To understand power, it is helpful first to consider an elementary testing situation in which there are only two candidate simple statistical hypotheses. The power of a test of one of those statistical hypotheses, given a fixed significance level, is the probability that the test will reject that hypothesis if it were false, i.e., if the other candidate hypothesis were true. When there is more than one simple alternative to the null hypothesis tested, the power is a function of whichever simple alternative is supposed as true.

What's important for present purposes is that replication measures based on NHST are insensitive to the power of either study being compared even though the power makes a difference to that study's reliability as a replication measure. How? Suppose that an initial study produces evidence for the existence of an effect. According to NHST, one of that study's results consists in the rejection of the null hypothesis. Even

[14] Semantic theories of vagueness are more popularly applied to the usual sorites cases philosophers analyze, in part because they take seriously the idea that vagueness is a semantical phenomenon of natural language. By contrast, here the goal of explicating "rejection" as a technical concept within the theory of NHST is not beholden to everyday linguistic activity. For similar reasons, ideal language approaches and those that take "rejection" as incoherent or vacuous are not apropos. (For more on these approaches to sorites paradoxes, see Hyde et al. (2018, §3) .)

[15] But see Douglas (2009, pp. 104–5) for some general considerations.

if the effect exists, i.e., the null hypothesis is false, a replication attempt with low power at that effect size will with high probability fail to replicate the initial study (Braver et al. 2014, p. 334; Simonsohn 2015, pp. 560–1). Conversely, if the original study did not reject the null hypothesis even though there is in truth a nonzero effect size, a replication attempt with low power at that effect size will with high probability replicate the initial study. In a word, replication measures based on NHST cannot reflect the reliability (or lack thereof) of the attempted replication to produce results that are not misleading.

Although ultimately I agree with the conclusion of this second criticism, too, the NHST advocate has grounds for defense similar to the ones for the previous criticism about the arbitrariness of the significance level. If one, again, takes the results of a study to consist solely in the *conclusion* about rejection that NHST offers, then what *kind* of study one runs (e.g., with high power) is irrelevant, as long as it employs NHST to arrive at that conclusion. In other words, the NHST advocate may accept that their replication measure is insensitive to a study's power but deny that this is relevant for replication if the power is not considered to be a part of or pertinent to the study's results.[16]

This defense, however, is in tension with the evidential goals of replication and classical statistics' endorsement of reliability's evidential relevance. The question of replication—whether one study produces "the same or sufficiently similar results as the original"—is not of scientific interest *intrinsically* but because it bears on the evidential and practical questions central to scientific inquiry: Which scientific hypotheses have been empirically established? Would one risk too much by using or building on a result to guide future inquiry, e.g., by presuming it true or empirically adequate? Degrees of evidence are the currency by which answers to these questions are purchased. In classical statistics, the degrees of evidence for or against a hypothesis depend not just on the results obtained, but the results that could have been obtained and with what probabilities, viz., the reliability of the testing procedure used (Fletcher and Mayo-Wilson 2021).

One way to maintain this defense would be to switch from the framework of classical statistical testing to a version of Bayesian statistics that denies the evidential relevance of reliability. For example, Etz and Vanderkerckhove (2016) have proposed replacing NHST replication measures, which depend on a significance level, with measures based on threshold values for Bayes factors. The Bayes factor is a comparative measure of the evidence that data provide, defined as the ratio of the likelihood of the data if one hypothesis were true to the likelihood of the data if the

[16] Simonsohn (2015) proposes another defense: calculate the effect size at which the original study has power of 0.33. Then, have the attempted replication test the hypothesis that the effect size is at least this large, or alternately check whether the attempted replication's confidence interval contains that effect size. Rejection (or the confidence interval's failure to contain the original point estimate) signifies a failure of replication. Simonsohn (2015, p. 565) requires that the replication have power of 0.8 at that effect size, which typically demands a sample size of about 2.5 times the original. Besides various ad hoc components, this proposal introduces two new parameters whose exact values are arbitrary and so only exacerbates the first problem with NHST. It also suffers from the asymmetry problem that befalls confidence interval-based measures of replication, which I describe in Section 6.2.

other, comparison hypothesis were true. For example, if the probability (or probability density) of the data according to one simple statistical hypothesis were 1/2, and that according to another were 1/10, then their corresponding Bayes factor would be $(1/2)/(1/10) = 5$, indicating moderate evidence for the first hypothesis over the second.[17] Bayes factors above one indicate some comparative evidence for the first hypothesis over the second, and vice versa for Bayes factors below one. (A Bayes factor equal to one indicates that the evidence is indifferently favorable). One can also compare non-simple statistical hypotheses—e.g., that the size of some effect is nonzero—by using the law of total probability, which requires calculating an integral that weights the likelihoods of the simple statistical hypothesis composing the non-simple ones by their prior probability.

With this extension, Etz and Vanderkerckhove (2016) compute the Bayes factor for the alternative hypothesis, that there is a nonzero effect size, against the null hypothesis of no effect. To do so, they assume a standard normal prior distribution for the effect sizes. They consider Bayes factors of at least ten to constitute "strong evidence" for the alternative hypothesis over the null. A replication attempt is successful when its results about strong evidence are the same as the original. Reanalyzing OSC's data, they found a replication rate of 75%, much larger than OSC's 39% using the NHST criterion.[18]

This way of defending a Bayesian version of an NHST replication measure still encounters problems of arbitrariness.[19] First, whenever there is more than one possible effect size, computation of the Bayes factor requires a prior over the simple hypotheses representing them. As Etz and Vanderkerckhove (2016, p. 3) admit, "Other analysts could reasonably choose different prior distributions when assessing these data, and it is possible they would come to different conclusions." Thus the conclusions of such an analysis are not (at least approximately) invariant under analysts' equally justified choices. Second, the Bayes factor threshold of ten for strong evidence inherits same problems as setting significance levels for NHST.

There is a certain underlying unity to the primary objections against employing replication measures based on NHST or analogues. In every case, one can understand the objections as revealing the inadequacy of NHST in capturing what the

[17] Thus, Bayes factors for simple statistical hypotheses are just likelihood ratios; they do not require any information about the prior probability for the hypotheses being compared. Consequently, likelihoodists, who focus on this ratio and eschew prior probabilities when it comes to statistical inference and evidence (Hacking 1965; Edwards 1972; Royall 1997), can adopt the same procedure. The second of my two criticisms in the subsequent paragraphs does not depend on these priors either, so it applies equally to the likelihoodist.

[18] Actually, Etz and Vanderkerckhove (2016) employ a more complicating weighting system, using what they call "mitigated" Bayes factors, based on different scenarios for publication bias, the phenomenon that the results of published studies are not representative of studies performed.

[19] There may be other Bayesian ways of construing an NHST replication measure. For instance, by using the techniques of *prior elicitation*, researchers could construct a justified prior representing the beliefs of a relevant scientist (or an average from a group of relevant scientists) as well as their preferences that determine a threshold for the Bayes factors. However, except in the simplest cases, these techniques themselves involve modeling choices and idealizations, variations on which can significantly alter the priors and preferences represented (Stefan et al. 2020). Thus it is not clear that using prior elicitation in practice avoids problems of arbitrariness.

results of a study are such that they could be sufficiently similar to the results of another.[20] A single study rarely establishes the existence or absence of a phenomenon, but rather allots degrees of evidence to each possibility. Rejection and non-rejection are too coarse of a division of epistemic consequences from the possible data to accommodate this. (This suggests already that one should resist adopting dichotomous replication measures, a conclusion I draw in Section 8). Moreover, those results depend on their reliable production; replication attempts should have a fighting chance to provide veritistic results about hypotheses. In sum, replication measures based solely on NHST encounter the problems they do because they reify a synecdoche, mistaking a particular technical goal in statistical studies for the broader scientific goal of producing and accumulating evidence.

5 Effect size comparison

5.1 Effect size comparison: description

The point estimate for a study represents the effect size that the data from that study best support (i.e., that with which the data is most statistically compatible). So, one way to compare the results of two studies about the same effect is by comparing their point estimates. However, due to the variability of the data, it is usually extremely unlikely that the point estimates from a study and an attempted replication thereof are exactly the same. If the space of effect sizes has a natural distance function on it, as it often does, one can compute the distance between two studies' point estimates. But, how close would they need to be for the original study to count as being replicated? How can such measures take into account the differing variances of their estimates, which manifest in differently sized confidence intervals? These differing variances entail that the same distances between point estimates for one pair of studies may not have the same implications for replication as another, as high variance should increase the tolerance for differing point estimates when evaluated for replication.

One way to overcome these difficulties is to run a statistical test using both point estimates known as a two-sample test. This significance test applies not to simple statistical hypotheses representing effect sizes simpliciter, but to those representing the *difference* in effect sizes of the populations from which the two studies are drawn. One runs a significance test of the null hypothesis that there is no difference in effect

[20] In addition to the criticisms I've described, Simonsohn (2015, p. 561) makes two further criticisms of measuring replication via NHST. Both amount to the fact that NHST does not depend on the similarity of the studies' estimates of effect size. For example, an original study with a large estimated effect size could be replicated "successfully," according to a NHST replication measure, by a study with a small estimated effect size. Like the criticisms I've described, a defender of NHST could claim that these criticisms beg the question because they presume different conceptions of what the results of a study are. But these criticisms are equally well explained in the unified way I have suggested: NHST does not adequately capture what the results of a study are, and so no viable replication measure can be based on it alone.

size. The attempted replication is successful if and only if this null hypothesis is not rejected.[21]

5.2 Effect size comparison: critique

A study's point estimate of an effect size ostends the effect size that the study best supports. Because effect sizes in general are fine-grained and quantitative, basing a replication measure on a statistical test comparing two from different studies might then prima facie plausibly avoid some of the issues with NHST. Such progress in the end is limited, however, and comes with new problems. In this subsection I describe three. The first two are essentially the same as problems that befell NHST: the arbitrariness of the significance level used in the test leading to a dichotomous result, and the unreliability of low power replications. The third is a new problem, but one related to the underlying unity of the objections described at the end of Section 4.2 about the nature and content of a scientific study's results: effect size comparisons are not sufficient to capture whether the results of two studies are the same, as a replication measure must.

For the first two problems, recall that an effect size comparison between two studies employs a significance test on the null hypothesis that the effect sizes of the populations from which the two studies were drawn were the same. Because this is an instance of NHST, just with a different null hypothesis, it encounters many of the same problems as recounted in Section 4.2. First, it needs to employ some significance level as a boundary between replication and non-replication, but in practice this boundary is arbitrary, conventional, or open to influence from objectionable bias. Second, a replication attempt with low power to detect a difference in population effect size from the original has a high probability of success, even if the original study found evidence for a non-existent effect. This can be especially likely when the original's estimated effect size was small (Simonsohn 2015, p. 561).

The third problem is that a test for significantly different effect sizes between two studies may not sufficiently bear on the question of whether the results of the studies are the same. The main reason is that a study's point estimate is only a proper part of its results, which also include the contours of CIs as various confidence levels, tests of auxiliary assumptions, etc. If the effect-size-difference test does not reject because the two studies' point estimates are similar, then there is not sufficiently strong evidence to conclude that the populations from which the two studies' data are drawn are different. But that does not preclude evidence for a difference coming

[21] An important qualification: Although OSC do use this method, they do not highlight it to describe a replication rate for any *particular* psychological effect. Instead, because OSC are interested in aggregate rates of replication in social and cognitive psychology, they compute *paired difference* significance tests (both *t* and Wilcoxon signed rank) that compare the estimated standardized effect sizes (in terms of correlation coefficients) found in nearly one hundred original studies with those estimated standardized effect sizes found in attempted replications of those studies. In a word, this test is of the hypothesis that there is no difference in the effect size for the *aggregate* of replication attempts in comparison with their paired originals. (Their test rejected this hypothesis, finding that the replication effect sizes were in aggregate smaller than the originals). However, the underlying idea in this application is quite analogous to that when applied to individual replication attempts.

from other results, e.g., low overlap of confidence intervals. Conversely, if the test does reject, then there is only evidence that the populations studied were different. The source of the rejection is underdetermined: it could be a replicability failure, or it could be heterogeneity in the population. (This is plausible if both studies' point estimates have the same sign but different magnitudes). This is compatible with the replication attempt being a candidate direct replication if that heterogeneity was not specified in the original study.[22]

6 Effect size within replication confidence interval

6.1 Effect size within replication confidence interval: description

Recall that confidence intervals can be understood as estimates of effect sizes that take into account, in the sense just described in Section 5, the variability of the data even when they arise from a context with a single true effect size. They offer a range of effect sizes statistically compatible with the data observed. It is then quite natural to ask, as a way of assessing replication, whether an attempted replication's confidence interval contains the effect size point estimate of the original. Alternatively, one can ask whether the original's confidence interval contains the attempted replication's effect size point estimate. (OSC employ only the former, presumably because not all the original studies that they attempted to replicate produced confidence intervals, but this is of no consequence for present purposes).

6.2 Effect size within replication confidence interval: critique

Determining whether the original point estimate of the effect size falls within an attempted replication's confidence interval (CI) may seem to ameliorate at least one of the problems with replication measures based on NHST, namely that they are not sensitive to certain fine-grained results of studies, such as the reported p-values for hypothesis tests (OSC 2015, p. 4).[23] However, analogous and other problems remain: the same arbitrariness, conventionality, or possibility for bias afflicting the selection of the significance level affects the selection of the confidence level, and replication attempts with low power still have a high probability of success, even if the original study produced misleading results. Regarding the latter problem, the reasons are essentially the same as before: studies with low power against hypotheses of interest will produce point estimates with high variance, so any CI constructed

[22] Simonsohn (2015, p. 561) suggests another related problem, that effect size comparisons answer the question of "whether the effect of interest is smaller than previously documented . . . rather than whether a detectable effect exists." But this problem implausibly presupposes that the point estimates are not a part of the results that must be sufficiently similar in a replication.

[23] There is also a Bayesian version of CIs, called credible intervals. However, the same problems arise for credible intervals as CIs because these problems depend on features of intervals common to both. (Cf. similar comments by Simonsohn (2015, p. 567) .)

from them will be wide. The wider the CI, the more likely it will encompass the original study's point estimate.

In addition to these, there is a pair of new problems with CI-based replication measures. Both afflict the above described version as well as the version that checks whether the original study's CI contains the attempted replication's point estimate. The first problem arises from a formal asymmetry between how the original study and attempted replication are treated by the measures: It's possible for this measure to determine that one study replicates another, but not vice versa. (This is even *likely* if their powers are substantially different). This is a problem if one thinks, reasonably, that the results of a study do not depend on the time it was conducted or published or whether other studies were conducted or published before or after (except perhaps insofar as the studies' covariates are temporal). For in this case, if two studies have the same or nearly the same results, then each should replicate the other. In other words, as a relation it is asymmetric, despite it being a formalization of the binary relation between studies of having "the same or sufficiently similar results," which is symmetric.

This asymmetry problem arises from two facts, that CIs in two studies can be of different sizes and that the point estimate values they respectively contain need not be in their centers. (The first is in essence the origin of the problem above about low power replication attempts still having a high probability of replication). As an example only of CIs of different sizes, one study might have a point estimate of 1 with a CI of [0, 2], while another might have a point estimate of 3 with a CI of [1, 5]. As an example only of asymmetric CIs, one study might have a point estimate of 1 with a CI of [0, 4], while another might have a point estimate of 3 with a CI of [2, 6] Whether the one replicates the other depends on which is regarded as the original study.

The second problem concerns an interpretational issue with CIs. As I discussed above, CIs are interval estimates of effect sizes; they are not directly predictions about future data. The reason for this is that effect sizes represent hypotheses about the magnitude of an effect, not the distribution of possible values of estimators. Even with an effect size fixed, there will be variability in the data resulting therefrom and hence in estimators based on them. Checking whether a point estimate from one study falls in the CI of another treats the results of the CI construction procedure as a fixed quantity, like an effect size, when in fact it too is constructed from data produced with variation.

On the basis of this second problem, Patil et al. (2016) advocate switching CIs for *prediction intervals*. Prediction intervals are designed specifically to ameliorate this problem. As with CIs, one first fixes a number between zero and one as the interval's confidence level. Then, a procedure for producing a prediction interval from data yields a range (e.g., an interval) of values of a statistic that may be produced in a future study. The range will contain the next value for that statistic with a probability equal to the set confidence level. Also, as with CIs, the size of the interval tends to covary with the confidence level, so that higher confidence levels yield larger prediction intervals and lower confidence levels yield smaller prediction intervals.[24]

[24] There is also a Bayesian version of a prediction interval, but as I described in the previous footnote, switching to Bayesian methods doesn't preclude any of the problems with interval-based replication measures.

In some cases, prediction intervals also solve the asymmetry problem. When the interval is constructed from a pivotal quantity, like with many methods for constructing CIs, the prediction interval based on one study for a statistic in another will have the same width as the prediction interval based on the other study for the one. The reason for this is that the size of the interval depends in a symmetric way on the variability of the estimators in both studies. For many of these pivotal quantities, the prediction interval for a new statistic is also guaranteed to be symmetric about the point estimate for that statistic. However, this is not generally the case (e.g., when the support of the statistic is not isomorphic to the real line). This means that adopting prediction intervals does not completely solve the asymmetry problem.

One way to attempt to mend the problem would be to ask about the conjunction or disjuction of the two questions about the point estimate of one study falling in the prediction interval of another. Although this solves the formal asymmetry problem, it re-introduces problems of interpretation. Both the disjunction and conjunction appear as ad hoc solutions that move the replication measure away from answering the question of whether two studies have the same or similar results. In any case, any interval-based measure of replication still faces the same sort of arbitrariness in the selected confidence level as measures based on NHST or effect size comparison did in Sections 4.2 and 5.2, respectively.

7 Meta-analytic summary

7.1 Meta-analytic summary: description

Since one of the primary goals of statistical analysis and inference in science is to quantify the evidence for and against hypotheses, one way to assess an attempted replication is to describe the ways in which it does and does not alter the evidential support for hypotheses of interest (viz., effect sizes). The extent to which the (perhaps tentative) conclusions that our total evidence suggests change after a replication attempt might therefore be used to measure the attempt's success. Methods for amalgamating the evidence for and against statistical hypotheses fall under the heading of *meta-analysis*, so called because meta-analytic techniques contribute to statistical analyses whose data are themselves statistical analyses, viz., individual studies.

One way to implement this idea "weights each study by the inverse of its [sample] variance and uses these weighted estimates of effect size to estimate cumulative evidence and precision of the effect" (OSC 2015, p. 5), although there are many ways to implement meta-analysis sensitive to the variety of studies and statistical techniques they use.[25] The weighted estimates provide a direct point estimate of the effect size based on both the original and attempted replication studies, which can be used to run a null hypothesis significance test for the effect and produce a confidence interval of effect sizes, just like with an individual study.

[25] See, for instance, Rosenthal (1991), Lipsey and Wilson (2001), and Ellis (2010), or Cumming (2013). In Section 7.2, I address the question of whether these many ways present a problem for meta-analytic measures of replication analogous to the problems of arbitrariness discussed in Sections 4–6.

What does one then do with such tests, estimates, and confidence intervals? One option is to submit these new results to the replication measures described in the previous three sections, focusing on NHST, the comparison of point estimates, or the comparison of point estimates with confidence intervals. Each of these, in a sense, constitutes a distinct measure of replication.

OSC focus on the first measure: for each pair of studies, they report the result of NHST based on the cumulative evidence that the two studies provide.[26] The replication is successful just in case the result of the test for the meta-analysis is the same as the result of the test for the original study. (Because, again, the original studies are typically published only when they reject the null hypothesis of no effect, a replication typically will be successful on this measure if and only if the meta-analysis also rejects this null hypothesis). What distinguishes this from the replication measures discussed in previous sections is that instead of comparing the original study with the replication attempt, it compares the original study with a meta-analysis that combines the evidence from the original and the new study. This can thus yield conclusions different from replicability measures based on NHST alone, for two studies—the first with statistically significant results and the second without—can, when combined, still reach statistical significance.

But meta-analytic approaches to replication need not adopt any of the previous replication measures. One can instead just recompute point estimates and confidence intervals after each new replication study, an approach that Braver et al. (2014, p. 334) call *continuously cumulating meta-analysis* (CCMA):[27]

> In CCMA, instead of misleadingly noting simply whether each replication attempt did or did not reach significance [in the manner of NHST-based replication measures], we *combine* the data from all the studies completed so far and compute various meta-analytic indexes to index the degree of confidence we can have that a bona fide phenomenon is being investigated. In other words, the individual effect sizes of the entirety of completed studies are pooled into a single estimate. ... The CCMA approach therefore shifts the question from whether or not a single study provided evidential weight for a phenomenon to the question of how well all studies conducted thus far support conclusions in regards to a phenomenon of interest.

In addition, CCMA facilitates analysis of whether the effect sizes that a group of studies finds is more or less heterogeneous than expected. The extent to which they are indicates that the studies may not be testing exactly the same phenomenon, hence that one should interpret the meta-analytic estimate of the intended phenomenon's

[26] OSC (2015, p. 4) were only able to employ this technique with 75 of the original 100 studies they examined because limitations in the reported statistics of the remaining 25 studies precluded the necessary meta-analytic calculations.

[27] Braver et al. (2014) work in psychology, but are not the first to suggest meta-analysis for their discipline. Schmidt (1996, 1992), for instance, has advocated it as a general methodology for "cumulative knowledge" in psychology and only more recently suggesting it as a partial solution for some of the problems of the replication crisis (Schmidt and Oh 2016).

effect size with caution before attempting extra modeling to account for possible sources of publication, selection, or reporting bias (Braver et al. 2014, pp. 336–7).

The process of testing for unexpected (hence, unmodeled) heterogeneity and bias is the same in meta-analysis as the process of testing for the misspecificaion of a statistical model in other contexts, such as regression (Mayo and Spanos 2004). In regression, one infers how one variable or type of variable (the dependent ones) depends functionally on another variable or type of variable (the independent ones), allowing for some random variation. A regression analysis of a data set often begins with a set of simple assumptions, such as that the dependent variables depend only linearly on the independent variables. A statistical "goodness-of-fit" test assesses evidence against this assumption; if violated, then the usual estimators for the regression function's coefficients will be biased. One then replaces the simple assumption of linearity with something more complex and contextually appropriate, and begins assessment of the model again. The situation with meta-analysis is the same: one often begins with a simple meta-analytic model, perform checks of that model (e.g., for the censoring mechanisms arising from publication bias discussed in Section 7.2), then adds complexity as necessary. See, e.g., Rosenthal (1991, Ch. 7), Ellis (2010, Ch. 6), or Schmidt and Hunter (2015, Ch. 13) for textbook treatments of how to detect and correct for various types of heterogeneity and bias in meta-analysis in this way.

Two further comments on the foregoing quotation are in order. First, Braver et al. (2014, p. 338) take dichotomous NHST-based replications measures to be misleading for reasons similar to some of those I offer in Section 4.2, namely that whereas adopting any particular significance level threshold for declaring results to be scientifically significant imposes a distinction in kind, any replication measure must in fact reflect that evidence comes in degrees. Second, one may wonder about why the "shift" in the replication question that CCMA promotes is legitimate. In Section 8, I describe this shift in more detail, endorsing it over the other means of employing meta-analysis that still accept that replication measures should produce dichotomous outcomes. Before this, I turn to rebutting various general criticisms of meta-analytic replication measures.

7.2 Meta-analytic summary: defense

The most prominent and repeated criticism of using meta-analytic methods as replication measures is that the original studies that they combine with their respective attempted replications "have inflated [estimates of] effect sizes due to publication, selection, reporting, or other biases" (OSC 2015, p. 5). "Publication bias" refers to the fact that the type of results in scientific studies correlates with those studies' publication (Romero 2019, p. 4): results providing evidence for the existence of a novel, flashy effect are positively correlated with publication, while results reporting no such evidence, especially direct replications, are negatively correlated (Fidler and Wilcox 2018, §2.2). It operates at the level of individual scientific studies, acting as a kind of "missing data" mechanism for the corpus of studies on a subject. Selection or reporting bias, by contrast, operates at the level of the sections of data that scientists collect and report and the statistical tests they run. For instance, consider again a scientist interested in the effect of different culturing conditions on the efficiency of

pluripotent stem cells to differentiate into certain other types of cells. They may test this efficiency for a variety of cell types, but report only on the types for which the treatment yielded a rejection of the null hypothesis of no effect. By selecting or reporting only some of the data they collected and statistical tests they ran, the scientist may raise their chances of publication but distort the evidence their data provides, e.g., by not correcting their p-values or confidence intervals for multiple testing effects.

A consequence of the existence and direction of all of these biases is that the sizes of effects, even when those effects exist, will be overestimated in the conventional research literature, and this overestimation will bias the conclusions of any meta-analysis of that literature that does not explicitly estimate and correct for them (as described in the previous subsection), even if there are many replications published in the literature (Schmidt and Oh 2016). For instance, in a recent study of 17 meta-analyses of effects in psychology compared with large-scale pre-registered studies of the same effects, Kvarven et al. (2020) found statistically significant differences between 12 (using, essentially, the effect size comparison method of Section 5 with a significance level of 0.005). The average ratio of standardized effect sizes of replications to meta-analysis across the 17 pairs was about 1/3. This is strong evidence in favor of the existence and direction of the biases supposed.

Before addressing this objection, it will be helpful to get clear on exactly which aspect of meta-analytic measures it bears. For this, there is a useful analogy with the concepts of accuracy and precision in metrology (JCGM 2012). In this context, "accuracy" refers to a measurement technique's ability to produce a result close to the truth, while "precision" refers to the technique's ability to produce relativity consistent results in repeated uses. What a replication criterion measures is analogous to precision, i.e., whether the results of studies are consistent under varying circumstances. Publication, selection, and reporting bias, on the other hand, affects a meta-analysis' accuracy, i.e., whether its results produced reflect the quantity they estimate.

This effect on accuracy means that biases *do* in general make a difference to the *probability* that a particular replication study will yield a particular result according to a replication measure. Thus, even though they do not affect token replication measure applications, they do affect the aptness of any replication measure type to fulfill its purpose: inform whether the results of certain scientific studies are reliable. In this sense, their effects parallel those of the statistical power of a replication study on replication measures discussed in Sections 4–6. But they are therefore a matter with which users of *any* replication measure should contend, not just meta-analytic measures. This is not just a *tu quoque* response, for meta-analytic measures are well-positioned in particular to incorporate these corrections through modeling, as described in the previous subsection. Indeed, because the biases in question are probabilistic (statistical) effects over the population of potential studies on a topic, it seems that meta-analytic procedures are *needed* for this correction, regardless of which replication measure is used.

If these biases affect all replication measures, why has this criticism been applied only to replication measures based on meta-analyses? I suggest that it's because the traditional role of meta-analysis is to report the total evidence about an effect reliably,

while no such precedent exists for the other replication measures. This may have led the scholarship on replication not to emphasize that publication, reporting, and selection bias are equally problems for *all* replication measures, *all* of which take the results of studies at face value and aim to provide assurance on the reliability of those results. What's particular about meta-analytic measures is that they can integrate with other meta-analytic tools to model and correct for these biases.

After the objections concerning publication, selection, and reporting bias are set aside, another problem remains for versions of replication measures that take the meta-analysis only as an input to one of the other replication measures I have discussed: they are vulnerable to many of the same (or strongly analogous) criticisms of those measures. For example, as I had described in the previous subsection, OSC employ meta-analysis to pool the data for each study pair, then calculate using NHST whether the pooled data entail rejecting the null hypothesis. Using NHST in this context is open to the objections against it that I raised in Section 4.2, in particular about the arbitrariness of the significance level and the lack of reliability for low power replication attempts. The same objections concerning arbitrariness and uneven reliability for effect size comparisons and confidence interval methods also carry over.[28]

CCMA, however, is not vulnerable to these criticisms. Recall from the previous subsection that it reports the meta-analytic point estimate of the effect size and a confidence interval thereof, perhaps at various significance levels, for each new replication study completed. This allows one to compare quantitatively the evolving best estimates of effect sizes as evidence accumulates—in particular, with those from the original study. It also allow one to assess whether the effect size estimates and confidence intervals are more or less heterogeneous than one would expect, which would provide evidence against the assumption that the studies are in fact measuring the same phenomenon. Using a variety of combinations of these methods, Braver et al. (2014) show using simulations with effect sizes and powers common in psychology that CCMA outperforms purely NHST-based measures of replication in coming to accurate and precise conclusions.

A third criticism pertaining even to CCMA concerns the latitude with which a meta-analyst may make choices in the steps of their analysis, such as in their inclusion and exclusion criteria for relevantly similar studies and in their techniques for determining the relative weighting of different studies. This latitude has even led to some philosophers to draw skeptical conclusions of meta-analysis's epistemological utility due to the unchecked possibility for the influence of bias (Stegenga 2011; Jukola 2015; Romero 2016). I agree, however, with the assessment of Holman (2019) that

[28]Other objections do not carry over. The third problem for effect size comparisons—that they do not plausibly test whether the results of two studies are the same (Section 5.2)—does not carry over because the comparison between the results of a previous study and that of a meta-analysis no longer aims to explicate such a comparison of sameness. Instead, it tests whether the addition of the results of a new study to one's total evidence changes how the total evidence bears on hypotheses of interest. The asymmetry problem for confidence interval-based measures (Section 6.2)—that as a relation between studies they are not symmetric, hence do not capture sameness of results—does not apply for similar reasons.

meta-analysts are continually and creatively improving their techniques to detect and correct for various types of bias, including sober recommendations of different techniques' effectiveness in different situations (van Aert et al. 2016; Carter et al. 2019). For instance, Braver et al. (2014) recognize that publication bias preferentially censors initial statistically insignificant studies, and perform simulations showing that CCMA does well to correct this bias if replication studies do not suffer the same bias. Moreover, Holman (2019) points out historical evidence that scientific communities do not ignore the latitude that conflicting meta-analyses seem to endorse, but through self-correction come to a consensus about which techniques are the right ones to use in which circumstances. Consensus on meta-analytic results and techniques suppresses the influence of the sort of bias under discussion and supports those results' objectivity. What's more, during transient periods of underdetermination between different meta-analytic techniques that seem at the time equally justified, meta-analysts can compare these discrete options—a so-called "sensitivity analysis"—to see if conclusions of interest are robust between them (Carter et al. 2019). This contrasts with the continuous options for, e.g., setting significance thresholds and the dependence of replication measures on their values, where the criteria for a successful sensitivity analysis are unclear.

8 Conclusions and implications: changing the question

CCMA is radical because it rejects a presupposition of the question about replication itself, namely that replication is a dichotomous phenomenon: one study either replicates another or it does not. Is this rejection legitimate or does it merely change the subject? There is good reason to believe that it is legitimate. First, we should allow for the possibility that the original question's presupposed binary possible answers were a needlessly simplified explication of the function of replication, one that distorts its scientific goals. Indeed, the cumulative growth of pure and applied science depends on replications providing sufficiently strong evidence of the reliability of phenomena studied, not whether this is established through answering a yes/no question (Schmidt 1996; 1992). Rarely will one study present sufficiently strong evidence to establish the magnitude of an effect; more rarely still will a single replication attempt so establish it if positive or debunk it if negative (OSC 2015, p. 6). One might explain why the replication measures discussed in Sections 3–6 were subject to such problems, many recurring, by observing that they assumed that replication must be measured dichotomously.[29] This is to adopt a misleading framework of how studies typically accumulate evidence, especially in disciplines like psychology, where studies with relatively low power are common and only have evidential strength when combined. Thus, in another sense, CCMA is conservative because it retains the essential

[29] It may be possible to delineate admissible measures of replication by starting with and defending this negative conclusion as an assumption instead. Investigating this possibility, however, must await another occasion.

function of replication, to amplify the evidence about an effect so that researchers may decide for themselves whether, and if so how, to build upon that effect in their further scientific pursuits. Doing this quantitatively allows researchers to calibrate their decisions better than qualitative, dichotomous replication measures do.

In liberating us from the coarse dichotomy of "success" and "failure," CCMA implicates discipline-wide studies of replicability, such as that of OSC (2015): instead of summarizing replicability with a percentage of successes, one can instead produce spaghetti plots of effect size estimates, CI bounds, and their incremental changes in standardized effect size over increasing numbers of replications. One can also quantify the variability of these changes using measures of dispersion and heterogeneity for effect size estimates over the collection of attempted replications. Based on these, studies that would have otherwise been classified as "failed replications" can be interpreted either as falling within the expected range of variation, or signal the need for further investigation to explain the unexpected difference. Neither the original nor the replication attempt can be deemed at fault, a priori. The conceptual adequacy of CCMA as a meta-analytic measure of replication thus allows us to better assess the extent of our evidence for the multitude of scientific hypotheses investigated.

When one does this for the range of studies under focus by OSC, one can observe the trends of narrower CIs centered on smaller effects sizes than reported in the original studies: there is still evidence for the existence of many effects on which the original studies focused, but with smaller magnitude. But in no way does it debunk the importance and magnitude of the replicability crisis, which has many causes and manifestations (Fidler and Wilcox 2018; Romero 2019). Meta-analytic techniques also do not address or ameliorate these causes, nor do they annul the pernicious effects of bias and questionable research practices without testing for and modeling them (Schmidt and Oh 2016). Interpreted uncritically and under these pernicious effects, simple meta-analyses can still produce highly misleading summaries of our total evidence, as the oft-repeated criticism of them discussed in Section 7 rightfully holds. But as I also discussed in that section, researchers are continuing to expand and delineate a toolkit of tests and modifications (van Aert et al. 2016; Carter et al. 2019), many of which have become standards in textbooks on the subject (Rosenthal 1991; Ellis 2010; Schmidt and Hunter 2015), similarly to how researchers have developed a vast toolkit of tests of and modifications to simple linear regression. As Holman (2019) has emphasized, meta-analysts have made and continue to make consistent progress in these developments and resolving which competing techniques are better justified for a given application.

Acknowledgements Thanks to Katie Creel, Dan Malinsky, Conor Mayo-Wilson, Tom Sterkenburg, Kino Zhao, and two reviewers for comments on a previous version.

Funding This essay was written in part with the support of a Visiting Fellowship at the University of Pittsburgh's Center for Philosophy of Science and a Single Semester Leave from the University of Minnesota, Twin Cities.

References

Anderson, C.J., Bahník, Š., Barnett-Cowan, M., Bosco, F.A., Chandler, J., Chartier, C.R., Cheung, F., Christopherson, C.D., Cordes, A., Cremata, E.J., Della Penna, N., Estel, V., Fedor, A., Fitneva, S.A., Frank, M.C., Grange, J.A., Hartshorne, J.K., Hasselman, F., Henninger, F., van der Hulst, M., Jonas, K.J., Lai, C.K., Levitan, C.A., Miller, J.K., Moore, K.S., Meixner, J.M., Munafò, M.R., Neijenhuijs, K.I., Nilsonne, G., Nosek, B.A., Plessow, F., Prenoveau, J.M., Ricker, A.A., Schmidt, K., Spies, J.R., Stieger, S., Strohminger, N., Sullivan, G.B., van Aert, R.C.M., van Assen, M.A.L.M., Vanpaemel, W., Vianello, M., Voracek, M., Zuni, K. (2016). Response to comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–c.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454.

Begley, C.G., & Ellis, L.M. (2012). Raise standards for preclinical cancer research: Drug development. *Nature*, *483*(7391), 531–533.

Braver, S.L., Thoemmes, F.J., Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333–342.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Carter, E.C., Schönbrodt, F.D., Gervais, W.M., Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144.

Cox, D.R., & Hinkley, D. (1974). *Theoretical statistics*. London: Chapman and Hall.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals and meta-analysis*. London: Routledge.

Douglas, H.E. (2009). *Science, policy and the value-free ideal*. Pittsburgh: University of Pittsburgh Press.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B.A., Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347.

Earp, B.D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621.

Edwards, A. (1972). *Likelihood*. Cambridge: Cambridge University Press.

Ellis, P.D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis and the interpretation of research results*. Cambridge: Cambridge University Press.

Etz, A., & Vanderkerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLOS ONE*, *11*(2), e0149794.

Fidler, F., & Wilcox, J. (2018). Reproducibility of scientific results. In Zalta, E.N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition.

Fletcher, S.C., & Mayo-Wilson, C. (2021). Evidence in classical statistics. Written for Routledge Handbook of Evidence, Maria Lasonen-Aarnio and Clayton Littlejohn, eds.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B.A., Johannesson, M., Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, *75*, 102117.

Gilbert, D.T., King, G., Pettigrew, S., Wilson, T.D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–b.

Graff, D. (2000). Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics*, *28*(1), 45–81.

Graff Fara, D. (2008). Profiling interest-relativity. *Analysis*, *68*(4), 326–35.

Hacking, I. (1965). *The logic of statistical inference*. Cambridge: Cambridge University Press.

Harlow, L.L., Mulaik, S.A., Steiger, J.H., editors (1997). *What if there were no significance tests?* Lawrence Erlbaum Associates.

Holman, B. (2019). In defense of meta-analysis. *Synthese*, *196*(8), 3189–3211.

Hyde, D., Raffman, D., Sorites paradox (2018). In Zalta, E.N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition.

Joint Committee for Guides in Metrology (JCGM) (2012). International vocabulary of metrology — Basic and general concepts and associated terms (VIM), 3rd edition. https://www.bipm.org/en/publications/guides/vim.html.

Jukola, S. (2015). Meta-analysis, ideals of objectivity, and the reliability of medical knowledge. *Science & Technology Studies*, *28*(3), 101–120.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526.

Klein, R.A., Ratliff, K.A., Vianello, M., Adams, R.B., Bahník, v., Bernstein, M.J., Bocian, K., Brandt, M.J., Brooks, B., Brumbaugh, C.C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W.E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E.M., Hasselman, F., Hicks, J.A., Hovermale, J.F., Hunt, S.J., Huntsinger, J.R., IJzerman, H., John, M.-S., Joy-Gaba, J.A., Barry Kappes, H., Krueger, L.E., Kurtz, J., Levitan, C.A., Mallett, R.K., Morris, W.L., Nelson, A.J., Nier, J.A., Packard, G., Pilati, R., Rutchick, A.M., Schmidt, K., Skorinko, J.L., Smith, R., Steiner, T.G., Storbeck, J., Van Swol, L.M., Thompson, D., van't Veer, A.E., Ann Vaughn, L., Vranka, M., Wichman, A.L., Woodzicka, J.A., Nosek, B.A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152.

Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, D.C: American Psychological Association.

Kvarven, A., Strømland, E., Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434.

Lakens, D., Adolfi, F.G., Albers, C.J., Anvari, F., Apps, M.A., Argamon, S.E., Baguley, T., Becker, R.B., Benning, S.D., Bradford, D.E., et al (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168.

Larrick, R.P., & Feiler, D.C. (2015). Expertise in decision making. In Keren, G., & Wu, G. (Eds.) *The Wiley Blackwell handbook of judgment and decision making* (pp. 696–721). West Sussex: Wiley.

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: SAGE.

Machery, E. (2020). What is a replication? *Philosophy of Science*, *87*(4), 545–567.

Mayo, D., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science*, *71*(5), 1007–1025.

McCloskey, D.N., & Ziliak, S.T. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice and lives*. Ann Arbor: University of Michigan Press.

Morrison, D., Henkel, R., editors. (1970). *The significance test controversy*. London: Aldine Publishing.

Nosek, B.A., & Errington, T.M. (2017). Reproducibility in cancer biology: Making sense of replications. *eLife*, *6*, e23383.

Nosek, B.A., & Errington, T.M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691.

Nozick, R. (1981). *Philosophical explanations*. Cambridge: Cambridge University Press.

Open Science Collaboration (OSC) (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), ac4716.

Patel, R., & Alahmad, A.J. (2016). Growth-factor reduced Matrigel source influences stem cell derived brain microvascular endothelial cell barrier properties. *Fluids Barriers CNS*, *13*(6), 1–7.

Patil, P., Peng, R.D., Leek, J.T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, *11*(4), 539–544.

Reiss, J., & Sprenger, J. (2017). Scientific objectivity. In Zalta, E.N. (Ed.) *The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab*. Stanford University, winter 2017 edition.

Romero, F. (2016). Can the behavioral sciences self-correct? A social epistemic study. *Studies in History and Philosophy of Science*, *60*, 55–69.

Romero, F. (2017). Novelty vs. replicability: Virtues and vices in the reward system of science. *Philosophy of Science*, *84*(5), 1031–1043.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass*, *14*(11), e12633.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Beverly Hills, CA,: Sage. Revised edition.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276–1284.

Royall, R. (1997). *Scientific evidence: a likelihood paradigm*. London: Chapman and Hall.

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *27*(10), 1173–1181.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115–129.

Schmidt, F.L., & Hunter, J.E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*, 3rd edn. Thousand Oaks, CA: Sage.

Schmidt, F.L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, *4*(1), 32–37.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100.

Shanteau, J. (1992). The psychology of experts: An alternative view. In Wright, G., & Bolger, F. (Eds.) *Expertise and decision support* (pp. 11–23). New York: Plenum Press.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569.

Stefan, A.M., Evans, N.J., Wagenmakers, E.-J. (2020). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*. https://doi.org/10.1037/met0000354.

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*(4), 497–507.

van Aert, R.C.M., Wicherts, J.M., van Assen, M.A.L.M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, *11*(5), 713–729.

Wolfers, J., & Zitzewitz, E. (2006). Interpreting prediction market prices as probabilities. Technical report, National Bureau of Economic Research.